Building a Crisis Management Term Resource for Social Media: The Case of Floods and Protests

Irina Temnikova[†*], **Andrea Varga**[§♯*], **and Dogan Biyikli**[‡]
[†] Qatar Computing Research Institute, Doha, Qatar
[§]Department of Computer Science, Babes-Bolyai University, Cluj-Napoca, Romania
[♯]Department of Computer Science, The University Of Sheffield, Sheffield, United Kingdom
[‡] Independent Consultant, London, United Kingdom
[†]itemnikova@qf.org.qa, [§] varga.andy@gmail.com, [‡] dogan@pericula.org

Abstract

Extracting information from social media is being currently exploited for a variety of tasks, including the recognition of emergency events in Twitter. This is done in order to supply Crisis Management agencies with additional crisis information. The existing approaches, however, mostly rely on geographic location and hashtags/keywords, obtained via a manual Twitter search. As we expect that Twitter crisis terminology would differ from existing crisis glossaries, we start collecting a specialized terminological resource to support this task. The aim of this resource is to contain sets of crisis-related Twitter terms which are the same for different instances of the same type of event. This article presents a preliminary investigation of the nature of terms used in four events of two crisis types, tests manual and automatic ways to collect these terms and comes up with an initial collection of terms for these two types of events. As contributions, a novel annotation schema is presented, along with important insights into the differences in annotations between different specialists, descriptive term statistics, and performance results of existing automatic terminology recognition approaches for this task.

Keywords: terminology extraction, crisis informatics, social media

1. Introduction

The aim of this article is to present a *terminological resource for Crisis Management* tailored to social media (currently to Twitter). The article also presents a preliminary empirical study of the terms employed by Twitter users while mentioning or discussing *emergency events*.

We define *emergencies* (or *emergency events*) as dangerous situations which can occur to individuals, institutions and countries and can lead to “a substantial loss of life, money, assets, and productivity” (Schneid and Collins, 2001).

The primary goal of the terminological resource will be to support automatic recognition of emergency-related tweets, their classification and extraction of event-related information. In this way it will assist emergency professionals in gathering the detailed information they need (Temnikova et al., 2013), before taking a response decision. Examples of details crisis managers need are: i) What type of event is it? (flood, fire, or earthquake) ii) In what phase the event is? (beginning, already running for a certain time, just finished) iii) What kind of and how many damages there are? iv) Who are the event participants? v) Are there any victims? vi) Are there any additional complications? (e.g. gas leak in case of fire).

The reason for developing a specialized resource for Twitter is that social media (and specifically Twitter) are starting to be increasingly used by journalists, crisis managers and crisis informaticians to detect emergency events and extract crisis-related information (Imran et al., 2013; Varga et al., 2013; Chowdhury et al., 2013). The reasons for that are (Blanchard et al., 2012):

1. *Witnesses/victims prefer communicating information to social media rather than to the official channels.*
2. Due to that, *new emergency information appears in social media incredibly faster*, compared to standard information channels.

A very good example is the Asiana flight crash[1], the first news about which was known to the public 30 seconds after the crash, by a tweet posted by a passenger boarding another flight. All official statements (from the Fire Department, Asiana Airlines, and the government), started appearing from 2 hours after.

Our terminological resource aims to address the current limitations of the approaches collecting emergency information from Twitter, which *rely mostly on geographical information and on pre-defined, manually collected lists of concrete-crisis-specific hashtags and keywords*, by this ignoring the complete linguistic picture of Twitter expressions, used to characterise emergency events.

We consider that extracting terminology from Twitter can be a challenging task as previous terminology recognition approaches relied on long, well-written, and well-formed documents.

Besides supporting emergency event detection and information extraction from Twitter, this resource will be useful for the following Natural Language Processing (NLP) applications: Ontology population, Named-Entity Recognition, Information Extraction, indexing for Information Retrieval, Text Summarisation and Question Answering. It will also contribute to the general linguistic knowledge of the Crisis Management domain and the language used in social media.

The main contributions of our paper are the following: i) An innovative term classification schema, ii) Annotation guidelines for this task, iii) Lists of terms characteristic

[*] These authors have contributed equally

[1] http://simpliflying.com/2013/asiana-airlines-crash-crisis-management-sfo/. Last accessed on September 3rd, 2013.

to floods and protests, iv) Insights into the differences in annotation between two different specialists, v) Linguistic insights into the nature of collected terms, vi) Results of the performance of standard terminology recognition approaches for this task.

2. Related Work

Besides the increasing number of processing Twitter NLP systems and approaches (e.g. TwitIE, Bontcheva et al. (2013)), workshops (LASM, SASM, LSM workshops[2]) and challenges (#Microposts2014[3]), mining information from social media to support Crisis Management has started to gain attention only very recently.

The few existing approaches usually tackle three tasks: 1) Automatic identification of crisis-related tweets (Imran et al., 2014; Temnikova et al., 2013; Varga et al., 2013; Cano et al., 2013; Ireson, 2009; Robinson et al., 2013), 2) Sub-classification of crisis tweets (Imran et al., 2014; Ireson, 2009; Chowdhury et al., 2013), and 3) Extraction of specific crisis-relevant information (Varga et al., 2013; Ireson, 2009; Imran et al., 2013).

The existing approaches employ mostly geo-location (Ireson, 2009; Kumar et al., 2011) (based on geographic location mentions in the text or Twitter information), manually collected lists of hashtags/keywords (Sakaki et al., 2010; Temnikova et al., 2013), or a combination of both (Robinson et al., 2013; Imran et al., 2014).

The most similar approach to ours, about to appear in ICWSM 2014, is the one of Olteanu et al. (2014). They collected a crisis lexicon from tweets discussing 6 different crisis events, two of which are floods. Their aim is to add crisis-specific terms to AIDR[4] classifiers to improve the recall of crisis-related tweets retrieval. They collect completely automatically the most discriminative uni- and bigrams for a crisis from tweets manually annotated as crisis-releavant by CrowdFlower[5] users. Their approach differs from ours as they do not manually annotate crisis terms, do not study the domain in advance, do not plan to use language technologies to augment the lexicon, and do not take in consideration the terms, important for crisis managers. They also aim to collect specific crisis-related terms, e.g. terms, relevant only to the Alberta flood in June 2013.

In addition, our work is also similar to Vieweg et al. (2010), as it provides important descriptive statistics and insights about crisis communications in Twitter.

3. Research Hypotheses

In this article we aim at running a preliminary investigation of the nature of crisis terms, their form, and part-of-speech (POS) tags they are characterized with. We also want to test how much existing specialized crisis glossaries cover the terms we are interested in, and how well state-of-the-art automatic terminology approaches perform on this task. Due to the complexity of the domain and our lack of knowledge about it, in this paper we aim to test the following hypotheses:

1. *Different types of emergency events are characterised by different terminology.* We consider that different types of emergency events are: fire, earthquake, terrorist attack, flood, etc. We test this hypothesis by extracting terminology from two different types of events (*floods* and *protests*) and comparing their terms and part-of-speech (POS) patterns.

2. *Different instances of the same type of event can be characterised with slightly different terminology and sets of terms.* Instances of the same type of event are, for example, two different floods, which occurred in different moments, in different locations, and thus involved different circumstantial elements (e.g. one a dam, another a rainfall). Although we assume that different instances of the same event will have many terms in common, we think that they should also have a certain amount of differing terms, representing the elements they do not have in common. We test this hypothesis and the extension of this phenomenon by analysing two events of each type and by comparing their terms.

3. *The terms, which describe emergency events in Twitter go beyond the standard notion of Noun Phrases (NP).* We want to test this hypothesis, in order to understand if the current terminology recognition and extraction approaches based on NP would work for this task. We test this hypothesis by investigating the part-of-speech patterns of annotated terms.

4. *The crisis terminology used in Twitter differs from the standard Crisis Management (CM) one*. We test this hypothesis, in order to motivate why we want to develop a terminological resource specifically for Twitter. We test this hypothesis by checking how many terms taken from existing specialized crisis management glossaries appear in our tweets and how many correspond to the terms we collected via manual annotation.

5. *Automatic state-of-the-art approaches perform badly for our task.* We want to test this hypothesis in order to determine whether we could use existing automatic approaches for our task, or we need to develop a specifically tailored one. We test this hypothesis by comparing the output of standard SoA terminology recognition (TR) methods with our manual annotations.

4. Methodology

Our approach for building the *terminological resource for Crisis Management*, can be summarised as follows:

1. Compilation of *emergency events* corpora.

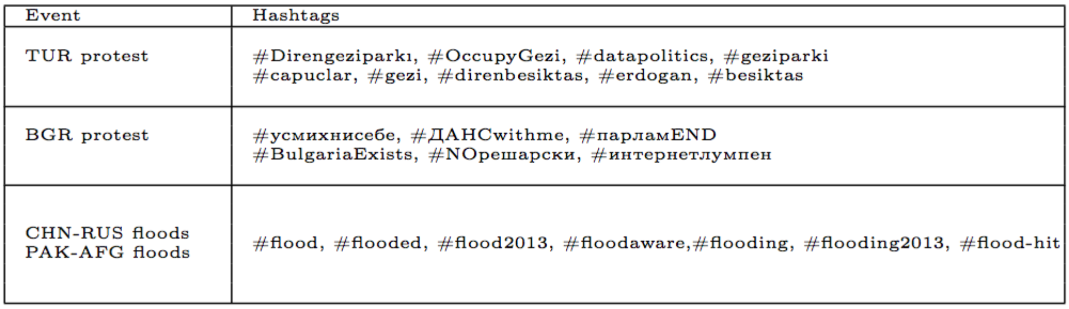
2. Annotation of emergency-specific terms by a linguist and an emergency management professional.

[2] http://www.site.uottawa.ca/∼diana/eacl2014-social-media-workshop.htm. Last accessed on March 22nd, 2014.

[3] http://www.scc.lancs.ac.uk/microposts2014/challenge/index.html. Last accessed on March 22nd, 2014.

[4] http://aidr.qcri.org. Last accessed on March 22nd, 2014.

[5] http://www.crowdflower.com



Event	Hashtags
TUR protest	#Direngeziparkı, #OccupyGezi, #datapolitics, #geziparki #capuclar, #gezi, #direnbesiktas, #erdogan, #besiktas
BGR protest	#усмихнисебе, #ДАНСwithme, #парламEND #BulgariaExists, #NOрешарски, #интернетлумпен
CHN-RUS floods PAK-AFG floods	#flood, #flooded, #flood2013, #floodaware,#flooding, #flooding2013, #flood-hit

Figure 1: Event-specific hashtags

3. Analysis of the obtained terms.
4. Evaluation of SoA TR methods and existing crisis management glossaries for this task.

The following subsections describe these steps in detail.

4.1. Compilation of Emergency Events Corpora

Floods and *protests* have been selected as: i) they are examples of a *natural and a human-made disasters*, and ii) they *occur with high frequency world-wide*.

In 2013 devastating floods have occurred in: Colorado, USA (September); Afghanistan and Pakistan (August); South-West China (July); Alberta, Canada (June); North India (June); Central and Northern Europe (May-June). In 2013 large anti-government protests (often involving injured and dead), happened, among others, in Greece, Bulgaria, Romania, Turkey, USA, Brazil, and are on-going in several Arab countries. Although protests are not considered proper crisis events, we also analyse them as they also involve severe damages, including injured and dead people. For our analysis, we have selected two recent events of each type: the Turkish Gezi park *protest* (TUR), which started on May 28th, 2013 initially to contest the urban development plan for Istanbul's Gezi park; the Bulgarian *protest* (BGR) against the Oresharski cabinet which started on June 14th, 2013; the heavy China-Russia (CHN-RUS) *floods*, which started on August 10th, 2013 and hit parts of Eastern Russia and North-eastern China; and the Pakistan-Afghanistan (PAK-AFG) *floods* which began on July 31st, 2013 with heavy rainfalls, causing widespread flash flooding, and receded on August 5th, 2013. Both protests are still continuing at the moment of publication of this paper.

We crawled Twitter using Twitter public search API[6] for tweets (TW) posted during the analysed *emergency events*, over a period of two months (between 1 July and 31 August 2013). From this collection we randomly selected a sample of 500 English-language tweets, containing at least one event-specific hashtag (see Figure 1). The tweets belonging to the two different floods were distinguished by their time-stamp.

In order to filter out English tweets, we used the TextCat[7] tool, which has been found to achieve best performance on tweets (Derczynski et al., 2013). We furthermore selected at least one tweet for each day, to guarantee a large coverage of the full event.

4.2. Corpus Annotation and Pre-Processing

The corpora have been annotated by two annotators – a *linguist* (*Annotator1*), specializing in *emergency management language* and an *emergency management specialist* (*Annotator2*). Annotators with different backgrounds were selected to *compare their views*. As an additional hypothesis we supposed that *Annotator2*'s views will have more weight, as it would reflect more the needs of the domain. For the research presented in this article, *Annotator1* has annotated 500 tweets for each of the four events (in total 2000 tweets) and *Annotator2* – 100 tweets for each event (in total 400 tweets). Due to this, all statistics presented in Tables 1, 2, 3, 4, 5, and in the word clouds are based only on the annotations of *Annotator1* (the linguist). The lists of all terms collected by both annotators, along with the statistics for *Annotator2*, are accesible online[8]. The annotation guidelines[9] provide description and examples regarding the *five tags* which need to be used during annotation. Compared to other emergency situations annotations schemas (Corvey et al., 2012; Imran et al., 2013), our approach is *novel as it captures the relative importance of terms and tweets' actionability*:

- actionTweet ($actTW$) – a tweet inviting for action, e.g. *"Government's rescue is too slow"*. or *"People are searching for drinking water"*.
- actionTerm ($actTR$)– the expression triggering the invitation for an action, e.g. *"rescue is too slow"* or *"searching for drinking water"*.
- high-importance-Term ($highTR$) – a crisis term, bearing the most important information (e.g. *"severe flooding"* in *"Severe flooding inundates hundreds of villages"*.)

[6] search.twitter.com

[7] http://textcat.sourceforge.net

[8] The lists of all terms annotated by both annotators can be found at https://sites.google.com/site/missandreavarga/resources/, last accessed on March 22nd, 2014.

[9] http://www.irinatemnikova.eu/Version-to-upload-Tweets-emergency-annotation-guidelines.pdf.

Event	#W	#AvgW	#TR				#actTW	%MW(>=2)	%MW(>=3)
			#highTR	#medTR	#lowTR	#actTR			
TUR	**1,658**	10.03 ± 6.84	183	128	99	31	29	37.90%	14.51%
BGR	1,399	10.64 ± 5.85	229	57	88	80	71	48.70%	16.23%
CHN-RUS	1,324	19.18 ± 6.98	**568**	**300**	102	75	43	65.00%	**28.45%**
PAK-AFG	1,138	19.97 ± 5.07	476	180	**109**	**223**	**101**	**67.80%**	27.37%

Table 1: General statistics about the 4 *emergency events* analysed (having 500 tweets (TW)/event) based on the annotations of *Annotator1* (linguist), following removal of hashtags and URLs, lowercasing and stemming each words (W). #W corresponds to the total number of unique words, #AvgW stands for the average number of words/TW; #TR denotes the total number of unique *TR* in a corpus, #actTW refers to the total number of unique *actTW* in a corpus, %MW($>= x$) stands for the percentage of MW expressions (word length at least x) out of the total number of unique TRs in a corpus.

- medium-importance-Term ($medTR$) – a crisis term, repeating the main information, or bearing less important information, which cannot be classified as “low” (*e.g. “inundates”*).
- low-importance-Term ($lowTR$) – a crisis term, adding circumstantial information. (e.g. *“hundreds of villages”*)

Based on these annotation guidelines, *Annotator1* used all 5 suggested tags, and *Annotator2* – only four of them (excluding the medium-importance terms). Table 1 (corpora statistics) shows that the *flood* corpora contain longer tweets (#AvgW, Column 3), more annotated terms (TRs, Columns 4-7), and a considerably higher number of *actTW* and *actTR* than the *protest* ones (Columns 5 and 7).

4.3. Terms Analysis

4.3.1. Differences between Annotators

The inter-annotator agreement between the two annotators was computed using the Lenient measure for the F-measure from the GATE Annotation Diff tool[10]. This measure captures the partial overlaps between the annotations of the two annotators (e.g. *”protesting”* and *”people protesting”*; *”protests”* and *”anti corruption protests”* are considered partially correct). As hypothesized, there are significant differences between *Annotator1*’s and the *Annotator2*’s annotations. The inter-annotator observed agreement showed relatively low results for *lowTR*: 29% for TUR, less than 10% for BGR, PAK-AFG and CHN-RUS, and 40% for the PAK-AFG *highTR*. The agreement was relatively high for the TUR (66%), BGR (54%) and CHN-RUS (82%) *highTR*. Some of the differences between *Annotator1* and *Annotator2* are: *Annotator2* annotated less terms than *Annotator1*. *Annotator2* annotates more *multi-words (MW)* expressions. This tendency is the most visible in the *highTR*, where in average, *Annotator1* has 17% more *single-word (SW)* expressions. E.g. *Annotator2*: *“a massive crowd now moves to parliament”* vs. *Annotator1*: *“massive crowd”*; *Annotator2*: *“in danger of devastating flood”* vs. *Annotator1*: *“devastating flood”*. This specificity can be explained by *MW* expressions bearing more complete crisis management information. Our findings show that MW expressions are 40-50% of the *protests* terms and over 60% of the *floods* terms (See Table 1).

4.3.2. Differences between Events

An analysis of the lexical variation between different instances of the same type of *emergency event* revealed that the two *protests* corpora share only 12 terms in common (including *“riot police”*, *“protests”*, *“government”*, *“protesters”*, and *“demonstrations”*). Some terms specific to TUR are *“teargas”*, *“court order”*, *“water cannon”*, *“civil war”*, while to BGR – *“anti-government protests”*, *“pro-gov’t protest”*, *“siege to the parliament”*. The *flood* corpora share a bit more common terms, 25 (including terms such as *“flood(ing)”*, *“disaster”*, *“kill”*, *“heavy rains”*, *“rivers”*, *“caused”*, *“rescue”*). Some terms specific to CHN-RUS are *“residents are evacuated”*, *“torrential rain”*, *“moodslide”*, and to PAK-AFG – *“army deployed”*, *“flood relief camp”*, *“rain triggers flooding”*. Analysis of the common terms between *floods* and *protests* have shown less than 10 terms in common (*“police”*, *“government”*, *“doctors”*, *“dead”*, *“army”*). The similarities between the two *flood* and the two *protest* list of terms can be seen in Figures 2, 3, 4, and 5. More concretely, Figures 2 and 3 show the word clouds for the *highTR*s and Figures 4 and 5 – for all the terms (low-, medium-, and high-importance included). It can be seen that the most frequent *highTR*s are the same in the two floods and in the two protests. However, an interesting difference can be noted between the most frequent terms from *all terms* in the two protests (Figure 4). All these results prove our first and second hypotheses, and demonstrate that different types of events exhibit different terminology, and that differences (even if smaller) can be found at event instances level.

4.3.3. Part-of-Speech Patterns

In terms of part-of-speech (POS) patterns[11], we have observed that there are differences between terms with different level of importance, between annotators for the same event, between different events of the same type, and between different types of events. The most common general trends are: 1) The most frequent *highTR* for all corpora are Nouns (N) and Noun Phrases (NP), e.g. *“protest”*, *“flood”*; 2) The *lowTR* for BGR are often Verbs (V) or Verb Phrases (VPs), e.g. *“moves”*, *“continue”*, *“fill the street”*; 3) TUR terms include more adjectives than BGR (e.g. *“civil war”*, *“automatic weapon”*); 4) PAK-AFG are characterised by a high number of N+N sequences, while CHN-RUS with Ad-

[10] http://gate.ac.uk/

[11] We obtained the POS tags by running a recent POS tagger tuned specifically for tweets (Derczynski et al., 2013)

Event	#N	#V	#Adj	#N			#V		
				highTR	medTR	lowTR	highTR	medTR	lowTR
TUR	2,879	1,569	**596**	179	110	**69**	39	33	35
BGR	3,225	1,133	528	299	49	59	29	22	18
CHN-RUS	3,495	**1,165**	585	619	**237**	64	**222**	**129**	66
PAK-AFG	**3,903**	1,068	471	**621**	184	58	110	60	**80**

Table 2: Statistics about the 4 events analysed based on the POS patterns for the annotations of *Annotator1* (linguist), #N denoting the total number of nouns identified, #V denoting the total number of verbs found, #Adj denoting the total number of adjectives identified.

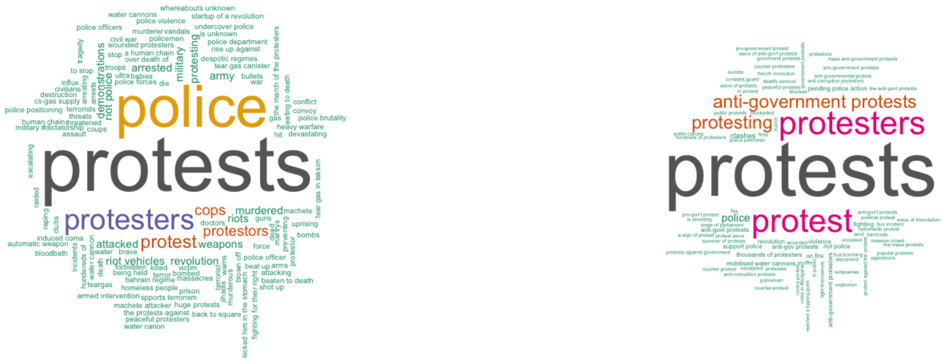


Figure 2: Word cloud for the protest events (TUR and BGR) for the high-importance-Terms (each term lower-cased) annotated by *Annotator1* (linguist).

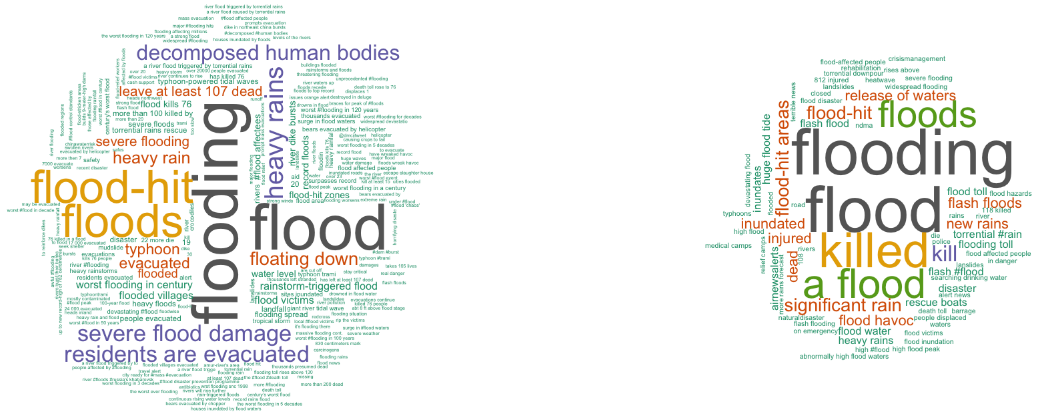


Figure 3: Word cloud for the flood events (CHI-RUS and PAK-AFG) for the high-importance-Terms (each term lower-cased) annotated by *Annotator1* (linguist).

jective (Adj)+N. Example: *"flood survivor"*, *"death toll"* (PAK-AFG) and *"strong winds"*, *"severe flood damage"* (CHN-RUS); 5) Finally, *Annotator2* terms often have unusual POS combinations, like: *"severe flooding inundates"* (Adj + N + V), *"have killed"* (V + V), *"flood toll rises"* (N + N + V). This variety confirms our third hypothesis.

4.3.4. Coverage of Specialized Glossaries

We also tested whether Twitter *CM terminology* differs from existing *specialized glossaries*, and thus collecting it is necessary instead of just searching for known terms. For this reason, we compared the annotations with *glossaries' terms*.

For the *flood* corpora a glossary of flood-related terms has been compiled[12]. For *protests*, we considered three glossaries: a glossary of riot terms[13], a glossary of non-violent actions terms[14], and a glossary of protest-related terms[15]. Very few glossary terms were found in our annotations: 33 terms (2.3% of the total glossary terms) have been identified in CHN-RUS (31(*highTR*)+2(*actTR*)+1(*medTR*)),

[12] http://www.fcd.maricopa.gov/Education/Glossary.aspx

[13] http://vm.uconn.edu/ pbaldwin/glosp5.html

[14] http://www.nonviolent-conflict.org/index.php/what-is-icnc/glossary-of-terms

[15] http://articles.chicagotribune.com/2012-05-18/news/ct-talk-nato-countdown-0518-20120518_1_nato-summit-pepper-spray-british-police

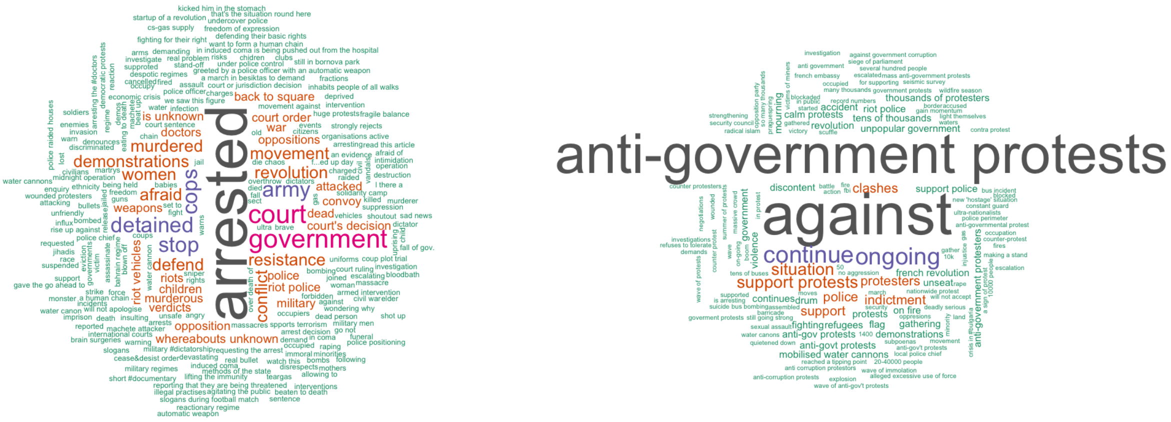


Figure 4: Word cloud for the protest events (TUR and BGR) for all the terms (each term lower-cased) annotated by *Annotator1* (linguist).

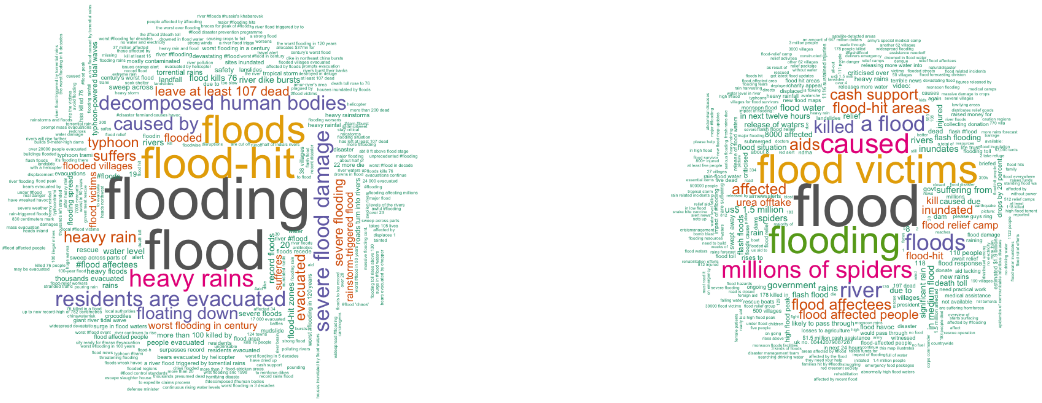


Figure 5: Word cloud for the flood events (CHI-RUS and PAK-AFG) for all the terms (each term lower-cased) annotated by *Annotator1* (linguist).

74 terms (2.3% of the total glossary terms) in the PAK-AFG *flood* corpus (72(*highTR*)+1(*medTR*)+1(*lowTR*)) (e.g. *"alert", "flood(ing)"*), 15 (2.1% of the total glossary terms) in the TUR corpus (8(*highTR*)+5(*medTR*)+2(*lowTR*)) (including *"resistance", "conflict", "weapons"*), and 32 terms (4.8% of the total glossary terms) (31(*highTR*)+1(*medTR*)) in the BGR corpus (including *"violence", "protest"*). This confirms our fourth hypothesis.

4.4. Automatic Evaluation

We also investigated the accuracy of automatic identification of emergency event specific *terms*, and *actionTweet* (actTW) tweets. As for the statistics, we used only the annotations of *Annotator1* (the linguist). Results about *Annotator2* will be available in the close future.

For the first task, we concatenated the annotations obtained for the *highTR*, *medTR* and *lowTR* terms together, and we employed three commonly used state-of-the-art TR tools for automatically extracting them: term frequency-inverse document frequency (TF-IDF), C-Value (Frantzi and Ananiadou, 1999), and TermRaider (Maynard et al., 2008). TF-IDF makes use of the term frequencies and inverse document frequencies, and balances the contribution of these two terms. As a result, a term achieves high TF-IDF value if the term has high frequency in one document, and low document frequency in the rest of the corpus. The other two approaches, on the other hand are hybrid, taking into account both term frequencies and contextual information about terms. C-value makes use of term frequencies and also examines their frequencies within nested terms. The main intuition here is that a term has high C-value if it has a high frequency and is not nested; or if it has been found nested in a small number of multi-word (nested) terms. TermRaider considers noun phrases as candidate terms as identified by a pre-processing tool, and then ranks them according to a scoring function. In this paper, we employ the Kyoto scoring (Bosma and Vossen, 2010), which considers the relationships among terms such as hyponyms and meronyms as connected in Wordnet to compute the final scores for the terms. In doing so, terms which have a high number of hyponyms and a high document frequency achieves a high Kyoto domain relevance score. These values are also normalised to lie between 0 and 100.

In the case of the first two approaches, we used the implementation available in the JATE term recognition toolkit

Event	Monthly			Weekly			Daily			Whole		
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
TUR	17.65%	9.27%	12.15%	17.8%	13.13%	15.11%	16.74%	14.67%	**15.64**%	21.15%	4.25%	7.07%
BGR	14.96%	11.52%	13.01%	11.56%	12.12%	11.83%	12.79%	16.97%	**14.58**%	24.44%	6.67%	10.48%
CHN-RUS	14.94%	3.27%	5.37%	14.29%	3.78%	5.98%	15.96%	8.56%	**11.15%**	17.02%	2.02%	3.60%
PAK-AFG	32.14%	5.34%	9.16%	25.81%	9.50%	13.88%	17.33%	11.57%	**13.88%**	40.00%	5.34%	9.42%

Table 3: Accuracy of emergency term recognition using TermRaider.

Event	Monthly			Weekly			Daily			Whole		
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
TUR	14.29%	0.39%	0.75%	11.11%	0.39%	0.75%	20.27%	17.37%	**18.71%**	9.24%	46.72%	15.43%
BGR	20.00%	3.03%	5.26%	20.00%	3.64%	6.15%	11.30%	15.76%	**13.16%**	5.99%	41.82%	10.49%
CHN-RUS	20.51%	2.02%	3.67%	26.32%	3.78%	6.61%	16.47%	10.33%	**12.69%**	8.18%	22.17%	11.95%
PAK-AFG	33.33%	7.72%	12.53%	31.25%	7.42%	11.99%	19.22%	14.54%	**16.55%**	8.59%	25.82%	12.89%

Table 4: Accuracy of emergency term recognition using TF-IDF.

Event	Monthly			Weekly			Daily			Whole		
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
TUR	9.22%	42.86%	15.17%	9.25%	39.77%	15.00%	9.41%	43.24%	**15.46%**	9.17%	43.63%	15.16%
BGR	5.87%	37.58%	10.16%	5.88%	36.97%	10.15%	6.03%	37.58%	10.39%	6.18%	40.61%	**10.73%**
CHN-RUS	7.83%	19.9%	11.24%	7.69%	19.4%	11.02%	7.67%	19.14%	10.95%	7.96%	20.4%	**11.45%**
PAK-AFG	8.49%	18.99%	11.73%	8.17%	22.85%	12.04%	8.12%	21.96%	11.86%	8.36%	24.04%	**12.40%**

Table 5: Accuracy of emergency term recognition using C-Value.

(Zhang et al., 2008), while for the latter approach we used the TermRaider plug-in with kyotoDomainRelevance in GATE (Maynard et al., 2008). We further applied a term frequency cut-off on the term importance scores, which we empirically set. In the case of the TermRaider approach, we set the term score cut-off to 41.00, while for TF-IDF and C-value we ignored terms with scores less than 0.01.

Given that traditional term recognition approaches have been designed for long document corpora, and that tweets are typically short, we have evaluated different pooling schemas for improving the performance of TR approaches on tweets. These pooling strategies aim to aggregate tweets into longer documents, which are more suitable for evaluating automatic TR models. The evaluated pooling strategies are as follows: *Monthly* pooling (pooling tweets posted in a given month), *Weekly* pooling (pooling tweets posted in a given week), *Daily* pooling (pooling tweets posted in a given day). We further refer to *Whole* for a document containing all the 500 tweets together for a specific event.

The results obtained for the different TR approaches and pooling strategies for the four analysed events are summarised in Table 3, Table 4 and Table 5.

As we can observe, in the majority of the cases the TF-IDF approach achieves the best overall results, except for the BGR, where TermRaider performed best. An explanation for this could be that TermRaider has identified a larger number of noun phrases as candidate terms for BGR, than for the other events. We can further notice that for these two approaches, the *Daily* pooling strategy performs consistently better than the other strategies, indicating that some emergency specific terms seem to be introduced on a daily basis. We can further notice that the different TR approaches are sensitive to the pooling strategy employed, especially considering the *Monthly*, *Weekly* and *Whole* strategies. For instance, TF-IDF and C-Value approaches favour longer documents (*Whole*) against the shorter *Monthly* and *Weekly* poolings. In the case of TermRaider approach, however, the *Weekly* and *Monthly* pooling strategies perform better than *Whole*.

One of the main drawback of the presented approaches was that they only identified NPs, ignoring VPs, which as shown in Table 2 constitute a big proportion of terms. They furthermore failed to recognise terms which were longer than 3 words long, which still cover a considerable percentage of terms. This indicates that a TR approach, which accurately identifies MW expressions and considers VPs would be better suitable for this task.

For the identification of $actTW$ types, a supervised SMO classifier from Weka[16] with Polynomial kernels was employed using 10-fold cross-validation. The tweets were first pre-processed: stemmed using Lovins stemmer, stopwords removed, lower-cased and only the top 1000 words kept. The results obtained on the different *emergency events* all showed relatively high results in *F1*, indicating that the task is relatively simple:

5. Conclusions

We have presented the first steps into collecting a *crisis management terminological resource*, reflecting the language used in Twitter based on a *novel importance- and actionability-based classification of terms*. Our experiments analysing two types of *emergency events*, and two

[16] http://www.cs.waikato.ac.nz/ml/weka/

Event	Precision	Recall	F1
TUR	90.6%	93%	**91.6%**
BGR	90.2%	90.3%	90.2 %
CHN-RUS	83.6%	84.2%	83.5%
PAK-AFG	79.6%	81.5%	80.3%

Table 6: Identification of actionable tweets.

instances of each type, proved all our research hypotheses and provided interesting linguistic insights, including that different *emergencies* are characterised by different terms and POS patterns. This poses difficulties for SoA TR approaches, which achieve poor performance on this task as they ignore multi-word expressions, and verb phrases. The tweets actionability task's results were however promising, achieving accuracy over 80%.

Our future work will consist in extending our analysis to other *crisis events* and investigate whether this classification of terms fits other events as well. Our annotation schema will be validated with a larger number of emergency professionals.

The resource will further be tested in the context of other tasks, such as automatic identification of crisis-related tweets, sub-classification of crisis tweets, template-based information extraction, and tweets ranking by importance.

6. Acknowledgements

The authors would like to thank Metodi Gerassimov for his suggestions regarding emergency journalism, Carlos Castillo and Alexandra Olteanu for providing us with details regarding the CrisisLex paper before it got published, and Muhammad Imran for allowing us access to the AIDR project.

7. References

Blanchard, H., Carvin, A., Whitaker, M. E., Fitzgerald, M., Harman, W., and Humphrey, B. (2012). The case for integrating crisis response with social media.

Bontcheva, K., Derczynski, L., Funk, A., Greenwood, M. A., Maynard, D., and Aswani, N. (2013). Twitie: An open-source information extraction pipeline for microblog text. In *RANLP*.

Bosma, W. and Vossen, P. (2010). Bootstrapping language neutral term extraction. In *LREC'10*.

Cano, A. E., Varga, A., Rowe, M., Ciravegna, F., and He, Y. (2013). Harnessing linked knowledge sources for topic classification in social media. In *ACM Hypertext 2014*.

Chowdhury, S. R., Amer-Yahia, S., Castillo, C., Imran, M., and Asghar, M. R. (2013). Tweet4act: Using incident-specific profiles for classifying crisis-related messages. In *ISCRAM*.

Corvey, W. J., Verma, S., Vieweg, S., Palmer, M., and Martin, J. H. (2012). Foundations of a multilayer annotation framework for twitter communications during crisis events. In *LREC*.

Derczynski, L., Ritter, A., Clark, S., and Bontcheva, K. (2013). Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In *RANLP*.

Frantzi, K. and Ananiadou, S. (1999). The c-value/nc-value domain independent method for multi-word term extraction.

Imran, M., Elbassuoni, S. M., Castillo, C., Diaz, F., and Meier, P. (2013). Extracting information nuggets from disaster-related messages in social media. *ISCRAM*.

Imran, M., Castillo, C., Lucas, J., Meier, P., and Vieweg, S. (2014). Aidr: Artificial intelligence for disaster response. In *WWW*.

Ireson, N. (2009). Local community situational awareness during an emergency. In *DEST'09*.

Kumar, S., Barbier, G., Abbasi, M. A., and Liu, H. (2011). Tweettracker: An analysis tool for humanitarian and disaster relief. In *ICWSM*.

Maynard, D., Li, Y., and Peters, W. (2008). Nlp techniques for term extraction and ontology population.

Olteanu, A., Castillo, C., Diaz, F., and Vieweg, S. (2014). Crisislex: A lexicon for collecting and filtering microblogged communications in crises. In *ICWSM 2014*.

Robinson, B., Power, R., and Cameron, M. (2013). An evidence based earthquake detector using twitter. In *Proceedings of the Workshop on Language Processing and Crisis Information*.

Sakaki, T., Okazaki, M., and Matsuo, Y. (2010). Earthquake shakes twitter users: real-time event detection by social sensors. In *WWW*.

Schneid, D. T. and Collins, L. (2001). Disaster management and preparedness.

Temnikova, I., Biyikli, D., and Boon, F. (2013). First steps towards implementing a sahana eden social media dashboard. In *SMERST*.

Varga, I., Sano, M., Torisawa, K., Hashimoto, C., Ohtake, K., Kawai, T., Oh, J.-H., and De Saeger, S. (2013). Aid is out there: Looking for help from tweets during a large scale disaster. In *ACL*.

Vieweg, S., Hughes, A. L., Starbird, K., and Palen, L. (2010). Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *SIGCHI*.

Zhang, Z., Iria, J., Brewster, C., and Ciravegna, F. (2008). A comparative evaluation of term recognition algorithms. In *LREC'08*.